1T-01 辞書の自動切り換え機能を考慮した翻訳辞書

羽鳥洋美　宮平知博
日本アイ・ビー・エム株式会社

1. はじめに

パターンベース翻訳システムPalmTree[1,2,3]に、辞書の自動切り換え機能[4]を実装した。これは、基本辞書よりも低い優先順位にある分野辞書のパターンデータをトリガにして、分野の優先順位を自動的に切り換えるというもので、訳質を向上させる上で重要な機能となっている。しかし、従来の翻訳辞書の構成では、辞書の自動切り換えによって期待される翻訳結果を得られない場合があり、翻訳辞書の構成を見直す必要があった。本論文では、この新しい辞書の構成と、辞書を再構成する過程でPalmTreeに実装した機能について述べる。

2. 従来の辞書の問題点

現在のPalmTreeでは、基本辞書とすべての分野辞書を翻訳に使用する。しかし、辞書の自動切り換え機能を実装する前のPalmTreeでは、基本辞書だけを常に翻訳に使用し、分野辞書はユーザーに指定されたときだけ使用した。このような使われ方を考慮して編集された従来の基本辞書と分野辞書には、各々次のような問題点がある。

2.1 基本辞書の問題点

基本辞書は、常に翻訳に使用される唯一の辞書であったので、本来は分野辞書に登録すべきデータでも、一般的によく使用される場合には、基本辞書に登録されていた。例えば、スポーツのチーム名や選手名、映画のタイトルなどの固有名詞である。これらは、辞書の自動切り換えのトリガとして有効なデータであるが、適切な分野辞書に登録されていないため、辞書の自動切り換えとは無関係になっていた。これは、翻訳中に期待する辞書に切り換わらないという問題を生じる。

2.2 分野辞書の問題点

分野辞書は、ユーザーに指定されたときだけ使用される辞書だったので、基本的な英語表現に分野特有の訳語が与えられたデータが数多く登録されている。例えば「ADV: this season = 今シーズン」というデータがスポーツ辞書に登録されている。「this season」は分野特有ではない一般的な表現なので、これは辞書の自動切り換えのトリガとして使用するには不適切なデータであるが、分野辞書のパターンデータなのでトリガとなってしまう。これは、翻訳中に予期せぬ辞書に切り換わるという問題を生じる。

3. 新しい辞書の構成

基本辞書の問題点は、データを適切な分野辞書に分類し直すことで解決できた。一方、分野辞書の問題点は、パターンデータを再構成することで解決できた。これについて3.1で述べる。さらに、訳質を向上させるために、分野を再構成した。これについて、3.2で述べる。

3.1 パターンデータの再構成

従来のパターンデータは1種類で、すべて辞書の自動切り換えのトリガとして使用されていた。しかし、スポーツ辞書にある「ADV: this season = 今シーズン」のような一般的な表現を含んだデータは、辞書の自動切り換えのトリガとしては不適切である。だが、その一方で、スポーツ分野の優先順位が上がったときにはこのデータを使用したい。そこで、新しい辞書では、分野辞書のパターンデータを「(p1) 辞書の自動切り換えのトリガになるパターン」と「(p2) 辞書の自動切り換えのトリガにはならないが、分野の優先順位が上がると使用されるパターン」の2種類に分類した。スポーツのチーム名や選手名、映画のタイトルなどの固有名詞は(p1)に、「ADV: this season = 今シーズン」のような一般的な表現を含んだデータは(p2)に分類される。これに対応して、PalmTreeでは、ある分野辞書が基本辞書よりも低い優

Dictionary System for Automatic Category Selection
Hiromi Hatori, Tomohiro Miyahira
IBM Japan, Ltd. (e-mail: hsaitoh@yamato.ibm.co.jp)
1623-14, Shimo-tsuruma, Yamato-shi, Kanagawa-ken

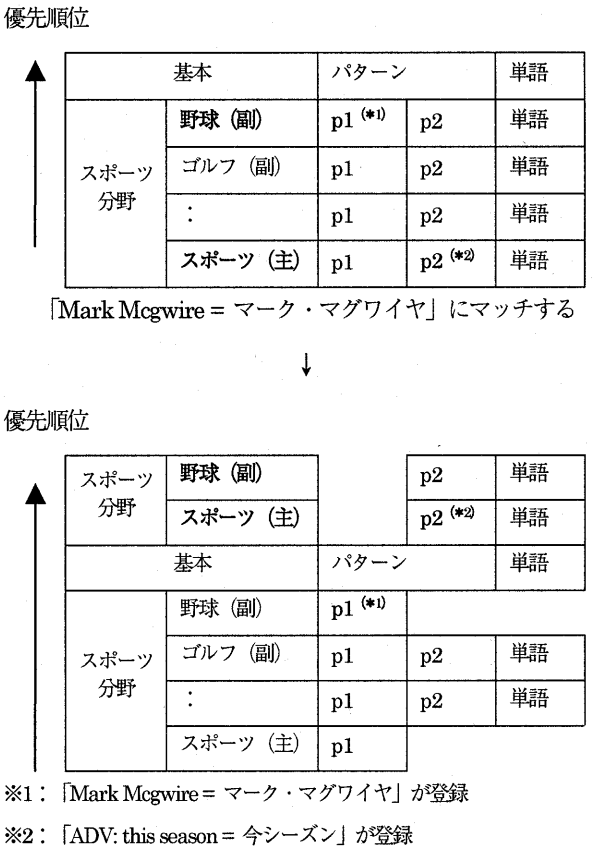
先順位にあるときには（p1）のみを使用し、その分野の優先順位が上がると（p2）も使用するように実装している。

3.2 分野の再構成

同じスポーツ分野であっても、サッカーでは「shot」を「シュート」と訳し、ゴルフでは「ショット」と訳したいように、従来の分野辞書には、より細かい分野に分類できるデータが存在する。その一方で、「ADV: this season = 今シーズン」のように、サッカーでもゴルフでも、同じスポーツ分野ならば共通に使用できるデータが存在する。そこで、新しい辞書では、分野を「主分野」と「副分野」という階層構造で再構成し、より細かい分野に分類できるデータは「副分野辞書」に、副分野間で共通に使用できるデータは「主分野辞書」に登録することにした。この階層構造を考慮して、PalmTreeでは、ある副分野辞書の優先順位が上がったら、その副分野辞書が属する主分野辞書の優先順位も同時に上げるように実装している。

3.3 翻訳例

スポーツ分野の副分野辞書である「野球辞書」の（p1）には「Mark Mcgwire = マーク・マグワイヤ」が、主分野辞書である「スポーツ辞書」の（p2）には「ADV: this season = 今シーズン」が登録されている。そして、初期状態では、「野球辞書」「スポーツ辞書」ともに、基本辞書よりも低い優先順位で使用されている。この状態で「People enjoy cherry trees this season.」という文章を翻訳すると、「人々は、この季節桜を楽しみます。」という翻訳結果が得られる。「this season」という表現が含まれているが、「ADV: this season = 今シーズン」というデータは基本辞書よりも低い優先順位では使用されないので、「this season」が「この季節」と翻訳される。一方、「Mark Mcgwire is a home run batter. He hit fifty home runs this season.」という文章を翻訳すると、「マーク・マグワイアは、ホームラン打者です。彼は、今シーズン５０本のホームランを打ちました。」という翻訳結果が得られる。これは、「Mark Mcgwire = マーク・マグワイヤ」によって副分野の「野球辞書」の優先順位が上がり、同時に「野球辞書」が属する主分野の「スポーツ辞書」の優先順位が上がるので、「ADV: this season = 今シーズン」が使用されて、「this season」が「今シーズン」と翻訳される。（図１）



※1：「Mark Mcgwire = マーク・マグワイヤ」が登録

※2：「ADV: this season = 今シーズン」が登録

【図１：新しい辞書での辞書の自動切り換え】

4. まとめ

PalmTreeの辞書の自動切り換えをより効果的に機能させるための新しい翻訳辞書の構成と、それに対応して新たにPalmTreeに実装した機能について述べた。今回の実装により、辞書の自動切り換え機能で、より適切な翻訳結果が得られるようになった。今後は、さらに適切な翻訳結果を得られるように、各辞書データの質と量を強化していく予定である。

参考文献

[1] Takeda, K., "Pattern-Based Context-Free Grammar for Machine Translation," Proc. of 34th ACL, pp. 144-151, 1996

[2] Takeda, K., "Pattern-Based Machine Translation," Proc. of 16th Coling, Vol. 2, pp. 1155-1158, 1996

[3] 渡辺, 武田, "パターンベース翻訳システム：PalmTree", 情報処理学会第55回全国大会, 1997

[4] 宮平, 神山, 羽鳥, "パターンベース翻訳システム PalmTree の訳語選択", 情報処理学会第59回全国大会, 1999